

平成30年度

修士論文

mRNA-seqを用いた  
Papaver somniferumの遺伝子発現解析

奈良先端科学技術大学院大学 情報科学研究科

野口 淳

指導教員： 小野 直亮 准教授

## mRNA-seq を用いた Papaver somniferum の遺伝子発現解析

ケシ (*Papaver somniferum*) は古くからモルヒネやコデインといったアルカロイド類を生成することが知られており、薬学的に重要な薬用植物である。そこで本研究では、この生合成経路を解明するために mRNA-seq を用いて一貫種のケシと一貫種を遺伝子改変したケシ、発現変動解析を行った。この解析により、変動する遺伝子および代謝経路を特定した。

キーワード： RNA-seq、次世代シーケンサー、*Papaver somniferum*

## master's thesis Overview Year 2018

### Gene expression analysis of Papaver somniferum using mRNA-seq

Poppers are known to produce alkaloids such as morphine and codin for a long time and are pharmacologically important medicinal plants. Therefore, in this study, in order to elucidate this biosynthetic pathway, poppy and consistent species of consistent species poppy and consistent species were genetically modified using mRNA - seq, poppy, expression variation analysis was performed. By this analysis variable genes and metabolic pathways were identified.

Keywords: RNA-seq, Papaver somniferum

# 目次

|      |                                                              |    |
|------|--------------------------------------------------------------|----|
| 第1章  | 序論                                                           | 1  |
| 1.1  | 背景                                                           | 1  |
| 1.2  | トランスクリプトーム解析                                                 | 1  |
| 1.3  | 代謝経路と代謝制御                                                    | 2  |
| 1.4  | アルカロイド                                                       | 3  |
| 1.5  | ケシ <i>Papaver somniferum</i> 及びオニゲシ <i>Papaver orientale</i> | 4  |
| 1.6  | 本論文の目的                                                       | 5  |
| 1.7  | 本論文の構成                                                       | 5  |
| 第2章  | 手法                                                           | 6  |
| 2.1  | 背景                                                           | 6  |
| 2.2  | 解析の流れ                                                        | 6  |
| 2.3  | シーケンサー                                                       | 6  |
| 2.4  | 配列アセンブル                                                      | 7  |
| 2.5  | de novo アセンブル                                                | 8  |
| 2.6  | アノテーション                                                      | 9  |
| 2.7  | マッピング                                                        | 10 |
| 2.8  | 発現変動遺伝子の検出                                                   | 10 |
| 2.9  | 2群間比較                                                        | 11 |
| 2.10 | パスウェイ解析による発現変動解析                                             | 11 |
| 第3章  | 結果と考察                                                        | 13 |
| 3.1  | m-RNAseq における発現量推定                                           | 13 |
| 3.2  | 発現遺伝子の検出                                                     | 13 |
| 3.3  | パスウェイ解析                                                      | 15 |
| 第4章  | 結論                                                           | 18 |
|      | 謝辞                                                           | 19 |
|      | 参考文献                                                         | 20 |

## 図 目 次

|     |                                      |    |
|-----|--------------------------------------|----|
| 1.1 | アルカロイド合成経路 . . . . .                 | 4  |
| 1.2 | ケシ . . . . .                         | 5  |
| 1.3 | オニゲシ . . . . .                       | 5  |
| 2.1 | 配列アセンブルの概要 . . . . .                 | 8  |
| 2.2 | de Bruijn グラフによるアセンブルの模式図 . . . . .  | 9  |
| 2.3 | 本研究の遺伝子発現解析の全体像 . . . . .            | 11 |
| 3.1 | PsIK と PsM を比較した M-A プロット . . . . .  | 14 |
| 3.2 | PsIK と POL2 を比較した M-A プロット . . . . . | 15 |

## 表 目 次

|     |                                                |    |
|-----|------------------------------------------------|----|
| 3.1 | PsIK と PsM 比較で発現量が増加した遺伝子が含まれるパスウェイ . . . . .  | 16 |
| 3.2 | PsIK と PsM 比較で発現量が減少した遺伝子が含まれるパスウェイ . . . . .  | 16 |
| 3.3 | PsIK と POL2 比較で発現量が増加した遺伝子が含まれるパスウェイ . . . . . | 16 |
| 3.4 | PsIK と POL2 比較で発現量が減少した遺伝子が含まれるパスウェイ . . . . . | 16 |

# 第1章 序論

## 1.1 背景

近年次世代シーケンサーや質量分析機器などの計測技術が急速に発展したことで一度に膨大な量の生物学的データである、遺伝子や蛋白質の構造・機能、発現やネットワークなどの生命情報を計測することが可能となった。それらは膨大な数のデータとして巨大なデータベースに蓄積されるようになり様々な生物情報が入手可能な時代となった。その代表的なものはゲノムの全塩基配列を解読するゲノムプロジェクトである。しかしながら、膨大な生命情報を獲得できたとしても、それらに何らかの意味を見出さなければならない。だが、配列を確認し手作業で生命情報に生物学的な意味を付与することは量的にも時間的にも困難であるといえる。

そこで登場したのがバイオインフォマティクス (bioinformatics) と呼ばれる学問分野である。生物学 (biology) に情報科学 (informatics) の技術を取り入れ、膨大な生命情報を獲得が可能になるとともに発展してきた学問である。バイオインフォマティクスの発展によって各種生物情報を情報学的な手法によって整理・解析を行うことで、知見の抽出が可能となった。バイオインフォマティクスにおけるコンピュータは人間に出来ないことをするのはではない。人間にも出来ることをより短時間に間違いなくこなすだけである。しかしながら、データの処理量が大きすぎて実際上は出来なかったことを可能にし、質的な差に結びつけ、それまでは得ることの出来なかった新たな知識の獲得が可能となり、分子生物学研究や創薬の分野において重要な役割を担うようになった。[1]

## 1.2 トランスクリプトーム解析

トランスクリプトームとは、特定の状況下における細胞中に存在する全ての mRNA の総体を指す呼称である。遺伝子はタンパク質の設計図であり、タンパク質は細胞内の様々な化学反応を含めた細胞の機能を規定する物質として、生命全般に関わる。すべての生物は固有の遺伝子セットを持ちながら、外部の環境の変化に適応するためや、細胞分化によってその役割を変化させるために、転写制御機構によって遺伝子の機能を調整している。ゲノム上の遺伝子は、mRNA に転写されタンパク質に翻訳されることによって機能する。これを遺伝子の発現とよび、遺伝子の発現は遺伝子の近辺に位置する転写制御領域に転写因子が結合することによって活性または抑制されることが知られている。そこで細胞内の mRNA の量を分析することで遺伝子の発現量を推定することを、トランスクリプトーム解析と呼ぶ。

トランスクリプトーム解析は以前までは DNA マイクロアレイを用いた手法が主流であったが、近年の技術発展によって次世代シーケンサを用いた手法が採用されるようになった。[2][3]

次世代シーケンサーは、細胞内の mRNA の配列を解読し、そこから遺伝子の発現量を推定する方法であり、これを一般的に RNA-Seq と呼ぶ。細胞から抽出した mRNA を短い断片にランダムに分割し、その両端にアダプターを付けることでライブラリを作成し、その後アダプター部位から配列の中央に向かって配列を読み進める。このような配列を読む操作をシーケンスと呼ぶ。

次世代シーケンサ登場以前はマイクロアレイを用いた遺伝子発現解析が一般的であった。マイクロアレイは、発現量を推定する際には配列に対応するプローブを作成する必要があるため、発現量を推定する際には配列が決定されている遺伝子しか扱うことができず、研究の発展しているモデル生物などの生物種に限定される。マイクロアレイと比較して次世代シーケンサーが持つ利点は、精度の高さと適応範囲の広さにあり、シーケンスされた配列の数によって決定され、測定できる発現量の上限がなく、それと同時に低発現の遺伝子も高感度で計測が可能である。また次世代シーケンサーは細胞内のトランスクリプトームをシーケンスするため、対象とする生物種のゲノム配列の情報を必要とせず、ゲノム配列が決定されておらず、個々の遺伝子の配列や機能の多くが未決定である非モデル生物にも適応することができるという利点がある。それに加え、発現量推定と同時に、一塩基型 (SNP) や突然変異などの個体間にまたがる塩基の違いを配列レベルで検出することができるのも次世代シーケンサーの大きな特徴である。しかしながら、次世代シーケンサーはトランスクリプトームをシーケンスすることから、解読された配列データから個々の遺伝子の発現量を変換するための計算的処理が必要となる。つまり解読した遺伝子配列がもとのゲノム上のどの遺伝子から特定する処理である。

次世代シーケンサが、シーケンシング可能な最大リード長は 100bp(ベースペア) 程度であるため、超音波破碎機などによって遺伝子の断片化を行い、その後に遺伝子断片のシーケンシングを行う。これにより mRNA-seq によって mRNA の断片配列が得られる。

### 1.3 代謝経路と代謝制御

代謝とは生命維持活動に必須のエネルギーの獲得や、成長に必要な有機材料を合成するために生体内で起こるすべてのせい化学反応の総称のことである。複雑な分子を単純な分子へ分解する過程でエネルギーと獲得する分解代謝 (異化) と、単純な分子から複雑な分子を構築するためにエネルギーの消費を伴う合成代謝である同化 (アナボリズム) がある。代謝によって生成されたエネルギーや化合物は、別の化学反応に利用される。これらの一連の反応の集合は代謝経路と呼ばれ、物質の変換経路として表される。

生物は代謝経路が適切に調整されることで維持されており、生体内における代謝は常に、外界の環境変動という形で影響を受け続ける。周囲の栄養素が枯渇すれば代謝経路の出発物質の取り込みを行うことができなくなり、環境中の温度が変化すれば細胞内の酵素活性が変化する。このような環境変動に対し細胞の駆動力としてのエネルギー獲得や、遺伝子の構成単位である拡散合成系のような一時代謝経路においては、外界の影響を受けにくい代謝経路が必要になる。反対に、代謝経路の下流に位置し、個別の機能に特化した二次代謝経路においては、特定の条件を引き金にして化合物の生成を促進したり抑制するような制御が必要となる。このような複雑な代謝制御を実現しているのが代謝経路の集合である代謝パスウェイと、化学反応の速度を規定している酵素である。

細胞内では数百種類の化学反応により、代謝活動が維持されていると考えられ、それらの化学反応の連鎖を生化学的な機能に基づいてグループ分けし、代謝ネットワークの部分集合



としてまとめたものを代謝パスウェイと呼ぶ。代謝パスウェイは直線、分岐、ループといった構造を持つものが多数存在し、複数の経路が一つに合流するなどして、これらが密に組み合わさり連動している。これによって一つの経路が分断されたとしても他の経路が働くことによって全体の機能が維持されることがある。代謝パスウェイは特定の機能を実現するために、様々なネットワークを構築しており、それらはおおまかな機能単位でクラスタを形成し、お互いに密に連携をすることで一つの巨大な代謝経路を形成して生命活動を支えている。

## 1.4 アルカロイド

アルカロイドは窒素原子を含む多くが塩基性を示す天然由来の有機化合物の総称で、その多くはアミノ酸から生合成され、細菌や植物、両生類などの様々な動物によって生合成されている。多くのアルカロイド防御と関連しており他の生物に対して有毒に働くが、中には薬理作用を示すものがあり、医療や娯楽として用いられているほか、毒性を利用して農薬として用いられているものも存在する。代表的なアルカロイドとしてコーヒーやお茶に含まれるカフェイン、煙草に含まれるニコチン等がある。

アルカロイドは複雑な骨格化合物を実験室で安価原料から合成することが難しい場合が多く、ケシなどの植物の抽出物などに供給に頼っているケースなども多いが、植物の生育には数ヶ月～数年を要することや植物体の乾燥重量の数パーセント程度しか含まれていないものも多く、供給に問題が発生している物質がある。

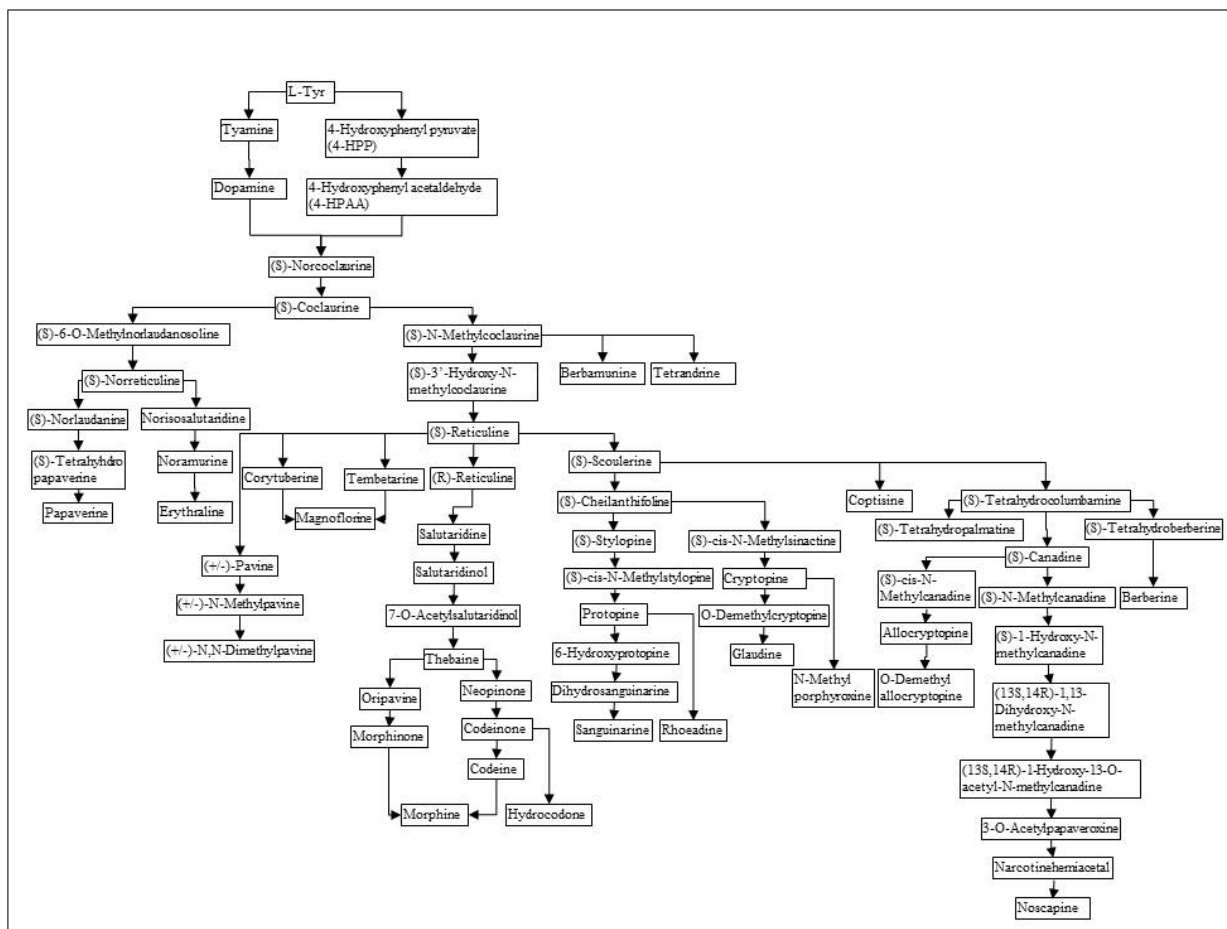


図 1.1: アルカロイド合成経路

## 1.5 ケシ *Papaver somniferum* 及びオニゲシ *Papaver orientale*

ケシ *Papaver somniferum* はケシ科ケシ属に属する1年草の植物で、日本ではアヘン法で栽培が原則禁止されてる種に指定されている。インドから小アジアにかけての西アジア原産とされており、高さ100~150cmで、葉は大きな長楕円形で、縁はギザギザになっており、5~6に直径10cmほどの白、赤、紫などの花を開く。栽培植物としての歴史は古く4大文明が興ったころにはすでに薬草として栽培されていたようであるが、その後麻薬用として利用されるようになり大航海時代を経て世界中に広まった。未熟果に傷をつけると乳液からアヘンが採取でき、それから精製されるモルヒネや、モルヒネを科学的に変異させたヘロインは麻薬に指定されている。

しかしながら、ケシは麻薬性鎮痛薬モルヒネおよびコデインの商業的供給源であり鎮痛鎮静剤として薬学的に用いられ、ガン患者の激痛を和らげるといったペインクリニックでの治療に不可欠であり、適切に使用されれば依存症に陥ることも少ない。[4]



図 1.2: ケシ

一方オニゲシ *Papaver orientale* はケシ科ケシ属に属する多年草の植物であるが、アルカロイドを生成しないことが知られており、高さ 50~70cm 葉が羽状に深く切れ込みがあり、白色、赤色、橙色、桃色などの大輪の花を咲かせる鑑賞植物として一般に販売もされている品種である。[6]



図 1.3: オニゲシ

## 1.6 本論文の目的

ケシはトリプトファン (L-tyr) を出発物質として 2 次代謝物としてアルカロイドである、パパベリン、ノスカピン、モルフィネなどを生合成することが知られているが、詳しいアルカロイド生合成機構はわかっていない。[5] そこで本研究では、アルカロイド生合成の基礎データを創出する目的で、ケシの発現プロファイル解析を行った。ケシと遺伝子改変によってアルカロイドの生成を阻害したケシ、ケシの近縁種であるがアルカロイドの生成を行わないオニゲシの遺伝子及び代謝の変動を調べることで、ケシのアルカロイド生合成機構の解明を目的とする。

## 1.7 本論文の構成

本論文は 4 章で構成され、1 章にて研究の背景及び目的、2 章では本研究で用いている主要な解析技術、3 章ではケシの遺伝子解析の結果及び考察について述べ、4 章にて本論文のまとめについて述べている。

## 第2章 手法

### 2.1 背景

全ての生物は、セントラルドグマと呼ばれる基礎原則で遺伝子情報を、伝達している。これは、DNA (=情報貯蔵物質) の情報が、RNA (=情報伝達物質) を介して、タンパク質 (=機能物質) の構造ならびにその構造によって生じる機能を規定している、という共通の原理によって生命活動を営んでいるという原理である。つまり、“生物は情報装置(情報伝達機関)であり、情報が物質の構造として維持、伝達される”というのが、現在の生物学における共通の認識である。この過程によって合成されたタンパク質が生体において成長促進や生命活動の維持などの様々な役割を担い、遺伝子情報が機能に変換される過程を遺伝子発現と呼ぶ。細胞内には遺伝子発現調節のためのいくつかの手段があるが、その大部分を担っているのがDNA から mRNA への転写量の調整であり、タンパク質の合成量を増加させる場合は転写量を増加させ、逆にタンパク質の合成量を減少させる場合は転写量を減少させる。つまり、mRNA の量を測定することで、どの遺伝子がどれほど発現しているかを推測することができる。これに着目するのが遺伝子発現解析であり、本章では遺伝子発現解析に用いる手法及びそれに関わる周辺技術について解説する。

### 2.2 解析の流れ

本研究では mRNA-seq を用いた遺伝子発現解析を行い、その後統計解析によって発現している遺伝子の検出を実行する。解析は

- (1) シーケンシング
  - (2) アセンブル
  - (3) 機能アノテーション
  - (4) マッピングによる発現量推定
  - (5) 発現変動遺伝子の検出
  - (6) パスウェイ解析
- という流れで行う。

### 2.3 シーケンサー

シーケンシングとはアデニン (A)、チミン (T)、グアニン (G) シトシン (C) の4種類の塩基から構成される塩基配列の並びを解析して決定を行うことで、これを行う機器のことをシーケンサーと呼ぶ。近年シーケンシング技術が次々と開発されたことにより、次世代シーケンサーが登場した。以前まではザンガー法が主流であり、高精度な配列読み取りが可能であったものの、大量の塩基対を処理するのに十分なものではなかった。そこでザンガー法以外のシーケンシング技術が次々と開発されたことにより、次世代シーケンサーが登場し

た。次世代シーケンサーはリードと呼ばれる短い塩基配列のシーケンシングを並列に行い、今迄のザンガー法を利用したキャピラリーシーケンサーと比べて性能が飛躍的に向上した。

次世代シーケンサーの登場により mRNA のシーケンシングも頻繁に行われるようになった。以前は DNA マイクロアレイを用いた遺伝子発現解析が一般的であったが、配列情報、機能情報が既知である場合には向くが、配列情報が未知の遺伝子に対する解析には向いておらず、ゲノム解読が進んでいる生物にしか適応できないという問題があった。一方で次世代シーケンサーはサンプルの細胞内に存在する mRNA の全情報を解析するため、遺伝子配列情報が既知である必要がなく、様々な生物種に対して遺伝子発現解析を行うことができる。DNA マイクロアレイを用いた遺伝子発現解析では発現量をシグナル強度として得るために精度が落ちてしまうが、次世代シーケンサーは mRNA の量を増やすほどに、高感度かつ高精度に発現量を推定することができる。現状の次世代シーケンサーでは、シーケンシングが可能な最大リード長は 1000bp 程度であるため、遺伝子の断片化を行い、その後遺伝子断片のシーケンシングを行っている。ゆえに mRNA-seq によって得られるものは mRNA の断片配列の集まりである。

## 2.4 配列アセンブル

mRNA-seq を用いた遺伝子発現解析では、既知の遺伝子配列をリファレンス配列として mRNA との比較を行い、配列類似性を認められたリファレンス配列に対して mRNA をマッピング、それらを数え上げることで各遺伝子の発現量を推定する。モデル生物などの研究が発展している生物種においては次世代シーケンサーによって得られた断片配列から新規に遺伝子配列を作成する必要がある。断片化された配列から元の配列を再構築することを配列アセンブリと呼び、その中でも、mRNA 断片から遺伝子配列を再構築することを de novo アセンブルという。[8] ある遺伝子から転写された mRNA の配列は一部を除いて同じ際列であり、それらを断片化したのちにシーケンシングを行うために、mRNA-seq によって得られた断片はそれぞれが互い違いに重なり合いを利用して、元の遺伝子の再構築を試みる。これによって再構築された配列を contig(コンティグ)と呼ぶ。

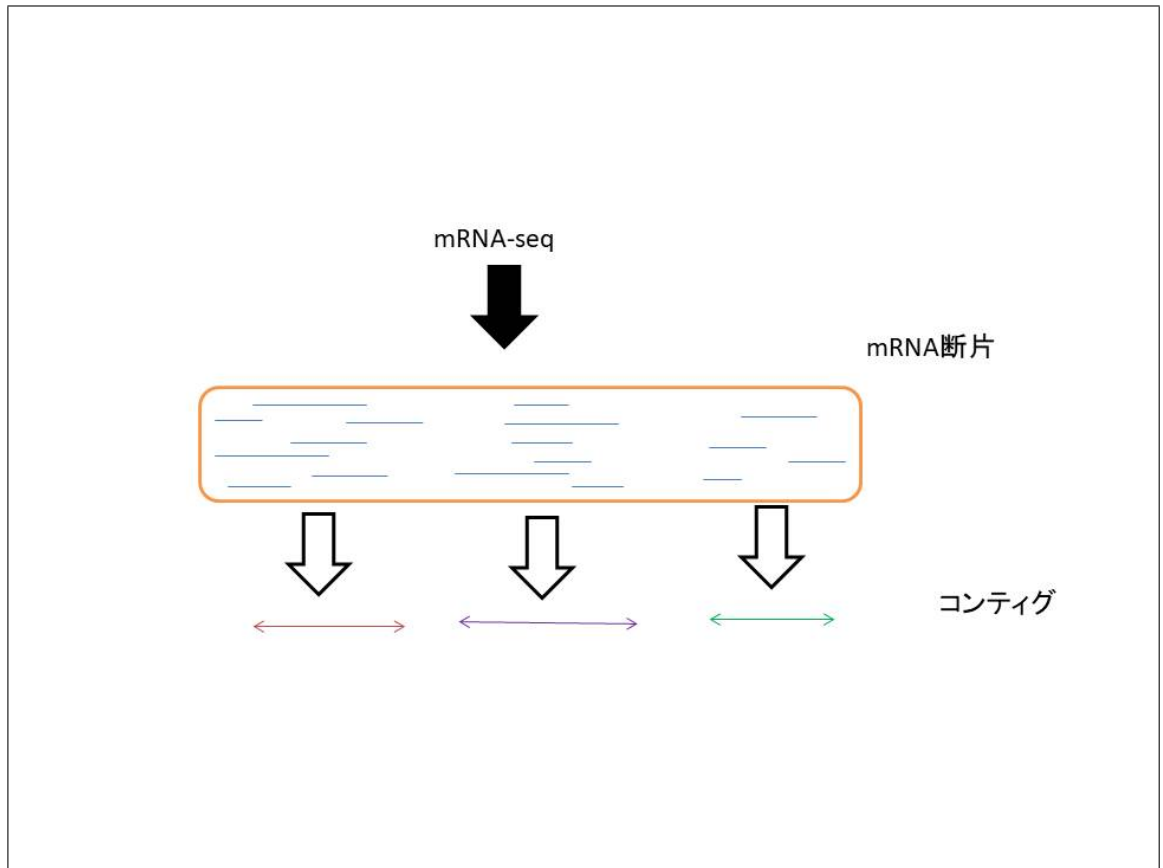


図 2.1: 配列アセンブルの概要

## 2.5 de novo アセンブル

配列アセンブルの中でも de novo 方式は、シーケンサから出力された配列の文字列をもとに、遺伝子配列を先に再構築する方法である。次世代シーケンサは遺伝子の断片を大量にシーケンスするが、配列が読み取られる場所は完全にランダムである。モデル生物は既にリファレンスとなるゲノム配列が解読され遺伝子のアノテーションがされている場合は、次世代シーケンサーで読まれた配列を既知のリファレンス配列と比較することで、その断片配列が何の遺伝子由来であるかを同定し、その遺伝子の発現量を定量することができる。一方で、非モデル生物などの参照可能なゲノム配列がない場合には、まず次世代シーケンサで読まれた配列の断片から元の mRNA 配列を同定することができるようになる必要がある。[9]

RNA-Seq の場合、mRNA の発現量が多ければその配列の様々な断片が繰り返しシーケンスされることになるため、お互いに重複した断片配列を多数得ることができる。それらの断片の重複した部分を手掛かりにつなぎ合わせることで、元の遺伝子の配列を復元することが可能となる。このように短い配列をつなぎ合わせて一つの長い配列を作ることをアセンブリと呼び、新規に遺伝子配列の再構築を行うことを de novo トランスクリプトームアセンブリと呼ぶ。[10]

いくつかの手法があるが今回は de Bruijn グラフによるアセンブルを使用した。シーケンサから得られた配列を、まず k-mer と呼ばれる k 文字の短い文字列の集合に分割する、

といったようにすべての場合で列挙し、k-mer を頂点として隣り合う頂点を辺でつなぎ、これをすべての配列について行ってグラフを作成し、同じ k-mer の文字を表す頂点を対応させることで複数のグラフを結合する。そして、グラフのすべての辺を通るようなパスを探し、その順番に頂点の配列をつなぎ合わせた結果が、アセンブルした配列となる。de Bruijin グラフはすべての辺を通るパスを考える、オイラーパス問題であり、効率的に解くアルゴリズムが存在する。そのため、配列の重なりを利用して単純につなぎ合わせることは回避できない重複遺伝子などの問題を解決することができる。[11]

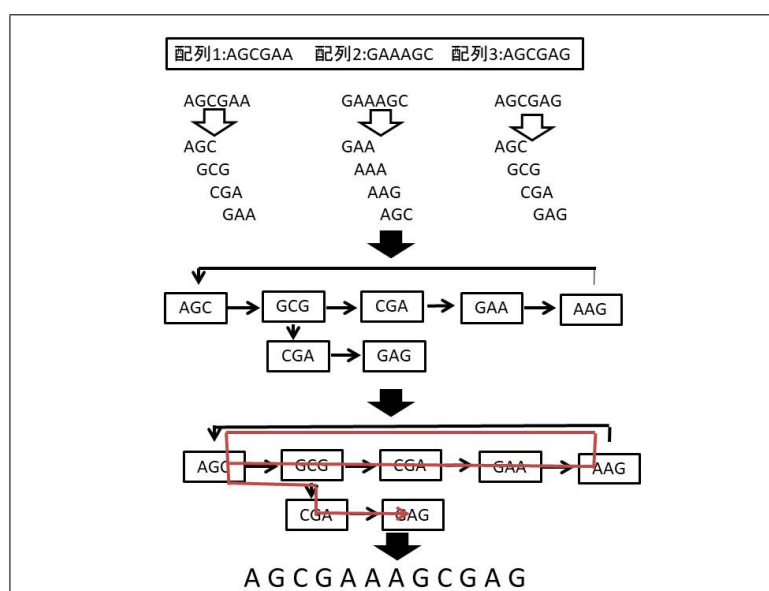


図 2.2: de Bruijin グラフによるアセンブルの模式図

配列 1(AGCGAA)、2(GAAAGC)、3(AGCGAG) が与えられた時、まず配列を k-mer に分割したのちに、隣り合う頂点同士を辺でつないで de Bruijin グラフを作成する。(この場合 k=3)。すべての辺を通るようなパスがパスウェイの結果として出力される。

## 2.6 アノテーション

De novo アセンブリによって得られたコンティグは単なる配列情報であるが、機能情報ははじめとする生物学的な意味を付加する必要がある、このように配列に生物学的な意味を付与することをアノテーションという。アノテーションには配列類似性に基づいたホモロジー検索が用いられる。ホモロジー検索は、手元にある配列と機能が既知である配列データベースとを比較し、データベース中から類似性が高い配列を検索し、この結果をコンティグにアノテーションを行うことができる。

DIAMOND とは、2つの配列間の相同性検索を高速に行うためのプログラムで、このプログラムは2つの配列の共通部分のみのアライメント (ローカルアライメント) を高速に行うことができる。

DIAMOND は、速度の 20,000 倍までの NR などのタンパク質参照データベースに対して DNA 読み取りまたはタンパク質配列を高感度で位置合わせ以前までのマッピングするための新しいプログラムである。[7]

## 2.7 マッピング

mRNA は DNA から転写されたものであり、遺伝子上には必ず転写元が存在していることになる。mRNA の塩基配列とリファレンス配列の塩基配列を比較することで mRNA の転写元を探し、対応付ける操作をマッピングと呼ぶ。マッピングは mRNA-seq によって取得した全ての mRNA 配列に対して大内、リファレンス配列のシーケンスアライメントを作成する。

そののちに、mRNA からリファレンス配列へのマッピングをしたのちに、そのマッピング結果をもとにして発現量の推定を行う。発現量の推定はリファレンス配列上に存在する各遺伝子にマッピングされた mRNA 配列の本数を数え上げるリードカウントと呼ばれる操作によって行われ遺伝子発現の量を定量化する際の指標として利用される。[12]

## 2.8 発現変動遺伝子の検出

より高い精度で発現変動遺伝子を検出するために、推定した発現量の TMM(Trimmed mean of M value) 正規化を行う。2 群間の発現量を比較すると、発現変動している遺伝子よりも発現変動していない遺伝子の数の方が多く、この時の 2 群間において、発現変動していない遺伝子の log fold-change(M 値) は 0 になることが期待されるが、そうになっていない場合がある。この状態で発現変動遺伝子の検出を行おうとすると、本来変動していないはずの遺伝子まで誤検出してしまう恐れがある。そこで TMM 正規化では以下のように計算を行う。[13]

(1) 全ての遺伝子に対して M 値及び A 値を算出する。

(2) M 値のうち上位 30

(3) A 値のうち上位 5

(4) 残った遺伝子の発現量から TMM 係数を算出

遺伝子  $g = 1, 2, 3, \dots, G$  において、M 値及び A 値は以下のように求められる

$$Mg = \log_2 \frac{Ygk}{Nk} - \log_2 \frac{Ygl}{Nl} \quad (2.1)$$

ここで、 $Ygk$  はライブラリー  $k$  における遺伝子  $g$  の発現量、 $Nk$  はライブラリー  $k$  におけるサイズを表している。また、TMM 係数は残った遺伝子をライブラリー  $r$  として求める。

$$Ag = \frac{1}{2} \left( \log_2 \frac{Ygk}{Nk} + \log_2 \frac{Ygl}{Nl} \right) \quad (2.2)$$



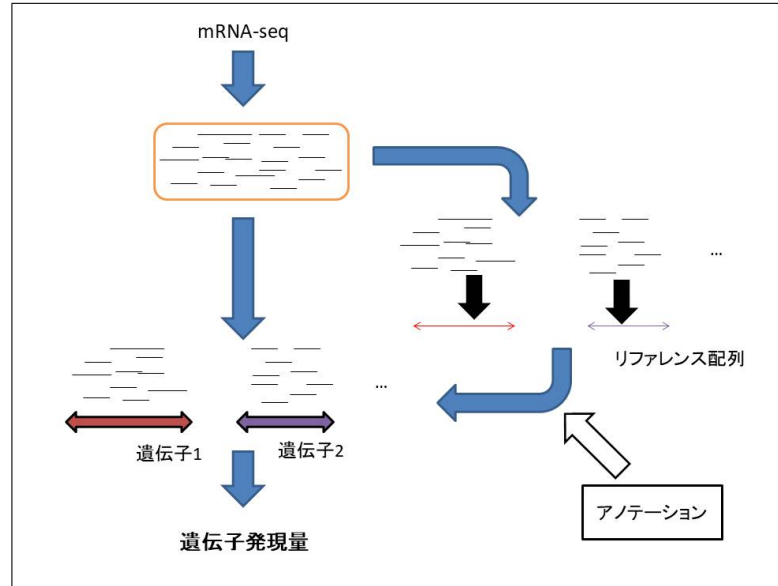


図 2.3: 本研究の遺伝子発現解析の全体像

## 2.9 2群間比較

2群間の遺伝子発現の有意差検定を行い、発現変動している遺伝子を検出する。遺伝子発言の分布は負の2項分布に従うことが示唆されているので、今回は負の2項分布を用いた有意差検定を行う。負の2項分布は確率  $p$  で成功するベルヌーイ試行を行った際、 $r$  回成功するまでに必要な試行回数の確率分布であり、以下の式で与えられる。

$$P(Y = y) = \binom{y + r - 1}{r} p^y (1 - p)^r \quad (2.3)$$

また負の2項分布の期待値及び分散は以下のように与えられる。

$$E(Y) = \frac{pr}{(1 - p)} \quad (2.4)$$

$$E(Y) = \frac{pr}{(1 - p)^2} \quad (2.5)$$

負の2項分布を用いた検定では、発現量が類似している遺伝子は発現量の分布も類似しているという考えに基づき分散を推定、上式から近似する分布を求める。そして、求めた確率分布に従った疑似乱数を発生させ  $p$ -value を算出する。また、多重検定を補正するために FDR(False Discovery Rate) を算出する。[14]

## 2.10 パスウェイ解析による発現変動解析

細胞は外部から取り込んだ物質をもとにして酵素による化学反応を繰り返すことで、エネルギーを獲得し、生命維持に必要な化合物に変換する。代謝によって生成されたエネルギーや化合物は、また別の化学反応に利用される。これら一連の反応は代謝経路と呼ばれ、

物質の変換経路として表される。この細胞の代謝経路において化学反応が効率よく進むには、一連の化学反応が連鎖して起こる必要がある。そのため、関連する酵素は共同して発現制御を受けている場合が多い。そのため反応の起点となる反応を触媒する酵素が発現しているときには、その下流の酵素も同時に発現している可能性が高いといえる。

生物の恒常性は、代謝経路における代謝物質の流れを適切に調節することによって維持される。生体内における代謝は、常に外界の環境変動という形で影響を受ける。周囲の栄養素が枯渇すれば代謝経路の出発物質を取り込むことができなくなり、環境中の温度が変化すれば細胞内の酵素活性が変化する。そのような外乱に対し、細胞の駆動力としてのエネルギー獲得や、遺伝子の構成単位である核酸合成系のような

ある 2 群間比較の遺伝子発現量データを検定し、遺伝子を発現変動遺伝子とそうでない非発現変動遺伝子を決定し、発現変動遺伝子にはどのようなパスウェイを持つ遺伝子が多いのかを調べることで、発現変動遺伝子が特定のパスウェイに集中していれば、何らかの生物学的な意味を持つのではないかと考えることができる。[15]

## 第3章 結果と考察

ケシはアルカロイド合成を行う重要な植物であるが、その生合成についての解析や代謝物の解析はほとんど研究されていない。そこで本研究では、mRNA-seq を用いた遺伝子発現解析を行い、どのような遺伝子発現の変動がみられるか研究を行った。

品種の差異や遺伝子変異によってどのような遺伝子の変動がみられるか解析を行った。

解析に用いたデータセットとして一貫種の PsIK をアルカロイドを生成するデータを 3 サンプル、一貫種のケシの遺伝子改変を行った PsM を 2 サンプル及びオニゲシの POL2 のデータを 2 サンプルをアルカロイドを生成しないをアルカロイドを生成するデータとした。

### 3.1 m-RNAseq における発現量推定

m-RNA における発現変動推定の詳細な流れは第二章に述べた通りである。はじめにケシの mRNA-seq データの de novo アセンブリを Trinity を用いて行い、その結果コンティグを得た。

de novo アセンブリによって得られたコンティグには機能情報が付与されていないため、Uniprot(The Universal Protein Resource) データベースに格納されている遺伝情報を、diamond を用いてアノテーションを行い結果として のコンティグがアノテーションされた。Uniprot は遺伝子の塩基配列やアミノ酸配列などから、タンパク質を構成するアミノ酸やタンパク質特性の収集された情報を機械的または手作業でキュレーションを行った情報が格納されている。アノテーションされたコンティグは e-value が 0.01 かつ、bit スコア が 1.0e-10 以下のものをアノテーション配列として用いた。

またアノテーションされたコンティグをリファレンス配列として mRNA-seq データをマッピングし、rsem を用いることで FPKM 遺伝子発現量の推定を行った。

### 3.2 発現遺伝子の検出

一貫種のケシ、遺伝子改変ケシ、オニゲシの有意差検定を行い、発現変動している遺伝子を検出する。遺伝子発現の分布が負の二項分布に従うと仮定し、二章で詳細を述べた通りに有意差検定を行った。

得られたコンティグに対して有意差検定を行うため、多重検定を補正するために FDR を算出し、FDR が 0.01 以下を満たしている遺伝子を発現変動しているものとみなした。また発現変動したとみなした遺伝子のうち logFC が 2.0 が以上のものは発現量が増加した遺伝子、logFC が -2.0 以下のものは発現量が減少した遺伝子とみなした。

有意差検定の結果、PsIK と比較して PsM の発現量が増加した遺伝子は 655、発現量が減少した遺伝子は 771 であり、PsIK と比較して POL2 の発現量が増加した遺伝子は 1219、発現量が減少した遺伝子は 1120 であった。図 3.1 は PsIK と PsM、図 3.2 は PsIK と

POL2 を比較した際の遺伝子の M-Aplot である。縦軸は M 値、横軸は A 値を表しており、赤で描画された点が発現変動遺伝子、黒で描画された点が発現変動なしの遺伝子を示している。

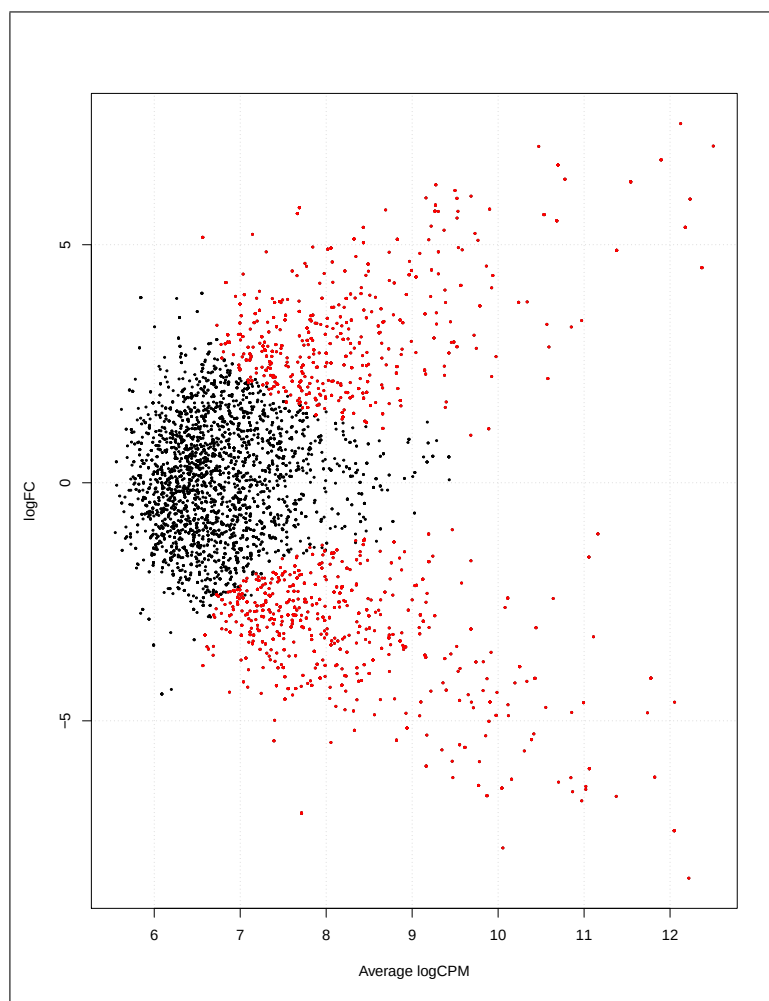


図 3.1: PsIK と PsM を比較した M-A プロット

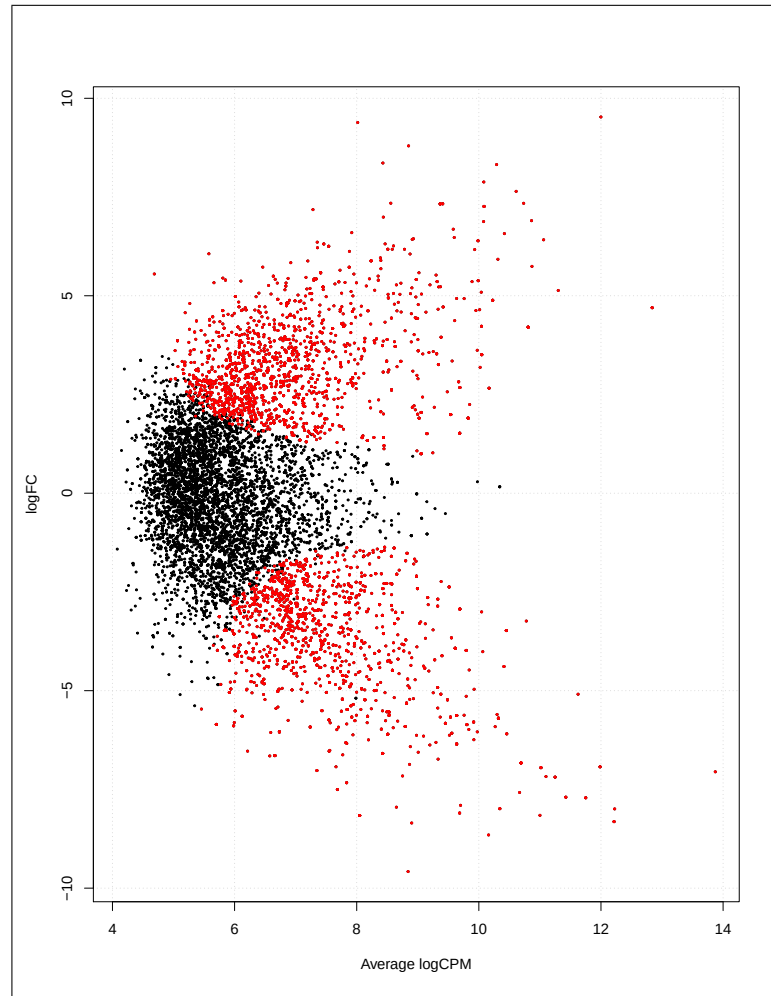


図 3.2: PsIK と POL2 を比較した M-A プロット

### 3.3 パスウェイ解析

これらの PsIK と比較して PsM と POL2 の遺伝子がデータベース上に存在する既知のパスウェイのうち、どのパスウェイに所属するかを調べた。この解析では遺伝子のうち KEGG ORTHROGY でアノテーションがなしているものに対し KEGGDB 上に格納されているパスウェイ情報にマップすることで特定のパスウェイに変動遺伝子が多く含まれているかを特定した。この解析では、遺伝子のうち KEGGDB 上に格納されているパスウェイ情報にマップすることで特定のパスウェイに変動遺伝子が多く含まれているかを特定した。

これらのパスウェイ解析の結果を考察する。PsIK と比較して PsM、POL2 の発現が増加していた遺伝子が含まれるパスウェイである Photosynthesis - antenna proteins、RNA degradation、Purine metabolism、Zeatin biosynthesis、Homologous recombination、Carbon metabolism の 6 つであった。これらは光合成に用いられる集光タンパク質、RNA 分解、プリン体代謝経路、プリン体・アデニンから誘導された植物ホルモンであるゼアチン、相同組み換え、炭素といったアルカロイド類の生産の有無にかかわらず変動している可能性が高いと考えられる。また、発現が減少していた遺伝子が含まれるパスウェイのなかでも Ribosome、

表 3.1: PsIK と PsM 比較で発現量が増加した遺伝子が含まれるパスウェイ

| パスウェイ名                            | KO pathway |
|-----------------------------------|------------|
| Photosynthesis - antenna proteins | ko00196    |
| RNA degradation                   | ko03018    |
| Purine metabolism                 | ko00230    |

表 3.2: PsIK と PsM 比較で発現量が減少した遺伝子が含まれるパスウェイ

| パスウェイ名                                   | KO pathway |
|------------------------------------------|------------|
| Zeatin biosynthesis                      | ko00908    |
| Photosynthesis - antenna proteins        | ko00196    |
| Sulfur relay system                      | ko04122    |
| Lysine biosynthesis                      | ko00300    |
| Arachidonic acid metabolism              | ko00590    |
| Pentose and glucuronate interconversions | ko00040    |
| Caffeine metabolism                      | ko00232    |

表 3.3: PsIK と POL2 比較で発現量が増加した遺伝子が含まれるパスウェイ

| パスウェイ名                            | KO pathway |
|-----------------------------------|------------|
| Photosynthesis - antenna proteins | ko00196    |
| Zeatin biosynthesis               | ko00908    |
| Homologous recombination          | ko03440    |
| Carbon metabolism                 | ko01200    |

表 3.4: PsIK と POL2 比較で発現量が減少した遺伝子が含まれるパスウェイ

| パスウェイ名                            | KO pathway |
|-----------------------------------|------------|
| Ribosome                          | ko00010    |
| Photosynthesis - antenna proteins | ko00196    |
| Photosynthesis                    | ko00195    |
| Fructose and mannose metabolism   | ko00051    |
| Phagosome                         | ko04145    |
| Oxidative phosphorylation         | ko00190    |

Photosynthesis - antenna proteins 、 Photosynthesis 、 Fructose and mannose metabolism Phagosome 、 Oxidative phosphorylation 、 Sulfur relay system 、 Lysine biosynthesis 、 Arachidonic acid metabolism 、 Pentose and glucuronate interconversions 、 Caffeine metabolism 、 は植物の生命活動において重要な役割をパスウェイのため、アルカロイド合成の有無にかかわらず発現が変動している可能性が高い。発現変動遺伝子がマップされたパスウェイの大部分はおそらく、アルカロイドの合成には関与せず絶えず変動しているものであると考えられる。現在所有するデータではこれらのパスウェイとアルカロイド合成の関連が推測できなかった。

PsIK と PsM を比較して発現が減少している遺伝子が多く含まれるパスウェイの中で興味深いのは Caffeine metabolism である。カフェインはアルカロイドの一種であり、プリン環であるキサンチン骨格に3つのメチル基がついた1,3,7-トリトリメチルキサンチンである。カフェイン生合成はキサントシンの合成とそれに続くメチル化の過程の2つから成り立っており、植物体内で、プリン環に N-7 位、N-3 位、N-1 位の順番でメチル基がつくことで合成される。

## 第4章 結論

本研究では、次世代シーケンサーを用いたトランスクリプトーム解析により、ケシのアルカロイド合成関連遺伝子発現を解析した。

KEGG データベースで公開されている既存の経路情報を利用して発現変動遺伝子のパスウェイ解析を行ったところ、アルカロイドを生産するケシー貫種とアルカロイドを生産しない遺伝子改変ケシを比較したところカフェインの生合成を行うパスウェイ遺伝子の発現が減少していることが解明された。

本研究では、目標であるケシのアルカロイド合成関連遺伝子の解明に至ったとは言えないが、今回示した結果により今後の解析に貢献をすることができたのではないと言える。これは、アルカロイド類の供給を植物からの抽出に依存しているために収量が不安定なケシの効率的な栽培のための様々な取り組みに対する足掛かりになると考えられる。

トランスクリプトーム解析によって対象とする生物のゲノム配列が利用できる時代になり、非モデル生物などのこれまで研究が発展していなかった分野においても大きな力を発揮することが期待される。大いに研究の幅が広がることを意味する。しかしながら、次世代シーケンサ登場以降、様々な種においてゲノム配列の決定が相次いで報告されているが、いまだにドラフトゲノムとしては未完成なものも多い。モデル生物であっても今なお完全な解読に向けては情報が更新されつづけており、非モデル生物で追従する場合、膨大な時間とコストを必要とするであろう。次世代シーケンサを用いたトランスクリプトーム解析は、ゲノム解読の代替案として非常に有効であるといえる。これからの生物学においてはトランスクリプトームのデータが蓄積されることにより、情報学的な側面から次世代シーケンサの解析手法を研究することは大きな重要性があるといえる。



## 謝辞

指導教官の金谷重彦教授には本研究室で研究を行う機会を与えて下さったことに、感謝いたします。指導教官の小野直准教授には、本研究の指導をはじめとして、研究室の活動や研究室のゼミナール副始動教員の MD.ALTAUF-UL-AMIN 先生および黄銘先生には、研究室のゼミナールにおいて多くの助言を頂いたことや指導して下さったことを、ここに感謝致します。副指導教員の笠原 正治教授にはゼミナールにおいて貴重な助言を頂いたことをここに感謝いたします。

また、日常の議論を通じて多くの知識や示唆を頂いた計算システムズ生物学研究室の皆様には感謝します。本論文において、多くの方々のご協力なしには完成しえなかったものであり、ご支援をして下さった多くの皆様に心より御礼を申し上げます。

## 参考文献

- [1] Michael L. Metzker."Sequencing technologies the next generation" Nature Reviews Genetics 11, 31-46 (2010) December 2009
- [2] Jay Shendure. "Next-generation DNA sequencing" Nature Biotechnology 26, 1135-1145 (2008) 09 October 2008
- [3] Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014
- [4] Weid M, Ziegler J, Kutchan TM "The roles of latex and the vascular bundle in morphine biosynthesis in the opium poppy,*Papaver somniferum*" PNAS 2004 September, 101 (38) 13957-13962, August 6, 2004
- [5] Thilo Winzer, Valeria Gazda, Zhesi He, Filip Kaminski, Marcelo Kern, Tony R. Larson, Yi Li1, Fergus Meade, Roxana Teodor, Fabian E. Vaistij, Carol Walker, Tim A. Bowser, Ian A. Graham, "A *Papaver somniferum* 10-Gene Cluster for Synthesis of the Anticancer Alkaloid Noscapine" Science 29 Jun 2012:Vol. 336, Issue 6089, pp. 1704-1708
- [6] J. D. Phillipson, A. Scutt, A. Baytop, N. Ozhatay ,G. Sariyar "Alkaloids from Turkish Samples of *Papaver orientale* and *P. pseudo-orientale*[1]" Planta Med 1981
- [7] Benjamin Buchfink, Chao Xie, Daniel H "Fast and sensitive protein alignment using DIAMOND" Fast and sensitive protein alignment using DIAMOND" Nature Methods 12, 59-60 (2015) doi:10.1038/nmeth.3176
- [8] Grabherr MG, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011, 29(7):644-52.
- [9] Haas BJ, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013, 8(8):1494-512.
- [10] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qian-dong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman "Full-length transcriptome assembly from RNA-Seq data without a reference genome" ,Nature Biotechnology volume 29, pages 644?652 (2011)doi:10.1038/nbt.1883

- [11] Pavel A. Pevzner, Haixu Tang and Michael S. Waterman, "An Eulerian path approach to DNA fragment assembly", PNAS 2001 August, 98 (17) 9748-9753.
- [12] Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015
- [13] Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013, 14:91.
- [14] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012
- [15] Huntley RP, Sawford T, Martin MJ, O'Donovan C. Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. Gigascience. 2014